Development and optimization of models to support clinical coding

Luís Filipe da Graça Fernandes Soares Baião

Department of Bioengineering, Instituto Superior Técnico, Lisbon, Portugal

November 30, 2015

Master Thesis Supervised by Professor Mónica Oliveira (IST) and Engineer Filipe Janela (Siemens)

Abstract

Diagnosis related groups (DRGs) correspond to a classification system of great importance to the financing of healthcare institutions. This classification system groups patients with similarities from the point of view of resources consumption and hospital stay duration, and is currently used in all hospitals in Portugal to define budget allocations for inpatient care to hospitals of the Portuguese National Health Service. This article explores the extent to which it is possible to predict DRGs with structured data from electronic health records (EHRs). Specifically, it develops a machine learning approach through the use of logistic regression models to predict DRGs, by using EHR structured data. The proposed approach was applied to a database with 5089 inpatient observations from the Fernando Fonseca Hospital. Furthermore, it was applied to the database a minimal-redundancy maximal-relevance filter in order to retrieve the most important attributes for each of the generated datasets. Each of the datasets was applied to a series of experiments in terms of finding the combination of variables that presented the best values for the sensitivity and predictive positive value performance measures. The results varied amongst predicted DRGs, with different model variations contributing to better model performance of distinct DRGs. For most DRGs, the best models presented high values for sensitivity and low values for predictive positive value. Results suggest that it is worth exploring the use of structure EHRs data to predict DRGs.

Keywords: Diagnosis Related Groups (DRGs); Electronic Health Records (EHRs); Portuguese National Health; Fernando Fonseca Hospital; Machine Learning; Structured Data; Minimal-Redundancy Maximal-Relevance Filter.

1. Introduction

Diagnosis-related groups (DRGs) consist in a classification system for ambulatory and admitted patients, and it divides them into clinically coherent groups according to resource consumption and duration of hospital stay. Presently, this classification system is applied at all Portuguese National Health Service (NHS) hospitals in Portugal, in order to budget allocations coming from the NHS regarding the hospitals' inpatient budgets [1]. For this reason, correct assignment of DRGs is an important issue in hospital management, as the incorrect classification of a DRG may have as a consequence a fine, if a DRG associated with a larger payment than the correct one is used, or a loss of income, if a DRG associated with a larger payment gets skipped [2].

Thus, the goal of this article is to explore the extent to which it is possible to predict DRGs through the use of structured data from electronic health records (EHRs) present in the information system of hospitals. For this purpose, it is intended the use of a machine learning algorithm, more precisely by means of a supervised learning approach, in which the information used for the prediction of DRGs consist in structured patient data present in the EHRs of the Fernando Fonseca Hospital (HFF).

This study contributes to current literature in DRGs prediction because, in contrast to other studies, it only relies on the structured data of the EHRs as source of information. Furthermore, this is also the first study, according to our information, which focus on the prediction of DRGs in the Portuguese context.

2. Background

Initially as a background to give a context to the problem at hand, it is given a perspective of the DRG classification system and its importance, in the Portuguese context. DRGs correspond to a classification system of admitted and ambulatory patients in clinically coherent groups with similarities from the point of view of resources consumption (costs of diagnosing and treating the patient) and duration of hospital stay [3]. They are constructed through the characteristics of the diagnosis and the patients' therapeutic profiles, in order to elucidate what each patient's consumption of resources in the hospital is [4].

Moreover, the necessary data for the DRG classification comes from general information that is available in hospitals, especially in discharge summaries. In fact, the designation of a DRG episode is completed through the use of the information registered in the discharge summaries and by using the following variables: main diagnosis in the form of international classification of diseases (ICD) code, secondary diagnosis (an ICD), performed procedures (an ICD), age, gender, discharge status and, in the case of newborns, the weight at birth [5]. This data enables an establishment with an average resource consumption rate to have estimations that can be used as a reference variable in asserting the duration of the patient's hospitalization [4].

Furthermore, DRG classification is a phased process, which procedure follows the ICD coding. After a medical record coder assigns diagnosis and procedure ICD codes in the hospital's medical records department, data concerning the ICD codes, along with the patient's age, gender, discharge status and even, in the case of newborns, the weight at birth, is sent electronically to an intermediary (coder), which inputs the data into a claim processing system [6]. This system is engineered to analyze all the cases and select the ones that require further revision before being classified into specific a DRG. Subsequently, through an automated algorithm (grouper), all cases are divided into 25 Major Diagnostic Categories (MDCs) and, after a follow-up process assigned into a specific DRG. Finally a file is submitted, containing all the payment data assigned to each DRG [7].

The current DRG system that is used in Portugal is a non-modified version of the All Patient DRG (AP-DRG) that was imported in 2006. This DRG system is applied to all NHS hospital and patients in Portugal, in the production lines of inpatient services, as well as surgical and medical ambulatory [1]. DRGs are presently used in DRG-based hospital budget allocations coming from the NHS, where the DRG-based hospital budgets come up to around 75-85 percent of the total of the hospitals inpatient budgets. [7]. Nonetheless the AP-DRG grouper is a property algorithm that is not available to use for research [8]. Thus, taking into account that a goal of this work is to develop a methodology to test the possibility of predicting DRGs based on structured data from EHRs, alternatives to the AP-DRG grouper should be searched in the available literature.

3. Literature Review

It was conducted a literature review of the diverse methods that can be used for the classification of DRGs, and in the prediction of DRGs. There were studied papers that used DRG groupers for the classification of DRGs in order to analyze their processes and source of information [9][10][11][12][13][14]. Furthermore, there were also studied articles in which machine learning techniques were used to predict DRGs or similar classification systems [15] [16].

It could be seen in this studies that, although using different methodologies, both the grouper classification methods and the machine learning prediction methods, used information that typically originated from patient records, as for both cases the type of information used tended to involve ICD diagnoses and procedure codes, and additionally demographic data.

Moreover, as seen by the meager number of studies that were found regarding the prediction of DRGs by machine learning methods [15] [16], this is a problem that still does not have many solutions according to the existing literature. Furthermore, it were not found any studies that contemplated the use of only structured data from the EHRs. For that reason it was proposed an overview of machine learning methods, in order to combine the already gathered information regarding the known studies, with a research about some of the most prominent types of machine learning algorithms, so as to develop a methodology in the prediction of DRGs by machine learning methods with the use of structured EHRs data.

In the overview of machine learning methods was given an insight in the field of machine learning, with emphasis on supervised learning and its most important algorithms which can be applied in the prediction of DRGs. Even though most of the studied methods could be applied to the prediction problem, they all presented different advantages and disadvantages amongst themselves. Thus, it was concluded that there was not a specific machine learning method that should be used in the prediction of DRGs, being the chosen method dependent on the desired analysis and possible limitations.

Since there were not found studies in the literature that use only structured data from EHRs to predict DRGs (as it was seen by the literature review), in the methodology section it will be suggested a new procedure type to predict DRGs. It will be proposed a machine learning methodology that uses structured data from EHRs to predict DRGs. The choice for the used machine learning algorithm will also be made in accordance to the advantages and disadvantages found in this review.

4. Methodology

Taking into account that in the methods available in the literature did not address all of the requirements defined for this study, it is suggested in this section a supervised machine learning methodology to predict DRGs based on information retrieved from structured fields of EHRs. The chosen machine learning method is the logistic regression that is used through the use of the software Stata® [17]. The process starts by the collection and treatment of structured data from the EHRs in order to obtain a database of attributes, to which a minimal-redundancy maximal-relevance (mRMR) filter is applied [18] that selects the variables in the database that relate the most with a specific chosen group of

DRGs (MDC 4 and MDCs 5+1+16 datasets). These chosen variables are then used to construct a dataset for the prediction of the previously selected group of DRGs, and later an assembly of different experiments in terms of classification models for the prediction of each DRG is proposed by using a logistic regression approach. Finally, the models in those experiments are evaluated in different ways by performance measures.

4.1 Collection and treatment of data

Initially in the methodology, it is necessary to retrieve the necessary and appropriated information to describe inpatient episodes from the EHRs, a process that is performed previous to this work [19]. This is achieved through the use of the data contained in the structural fields of EHRs in the Soarian EHR system, in which the information can be referred to as demographic data, clinical information or medical services and therapies. In the demographic group, the data comprises data blocks regarding sex, age, origin and discharge status. While in the medical services and therapies group the information consists of a data block related to (treatment) prescriptions and of another with patients' medication. Finally, in the clinical information group, the elements of the EHRs consist of data blocks related to patients' personal histories, allergies, diagnosis (ICD codes) and assessments.

To characterize the demographic information in the EHRs, several variables are taken into account, such as the sex, the origin and the discharge status, which are encode into categorical variables, and the variable of age, which is encoded through the use of its original values. Meanwhile, in the medical services, as well as in and therapy groups and the clinical information group, most of the data blocks are completely structured (the assessments data block is the exception), since they are composed by items that can be chosen from catalogs. Thus a binary variable can be defined for every component (item) of the catalogs, in order to characterize the absence or presence of a component of the catalogs in a given patient episode.

Oppositely, in the clinical information group, the assessments data block consists of both structured and unstructured data regarding medical concepts, from which only the structured data is considered (the unstructured data is discarded). Nonetheless, due to its diversity of values, nature (categorical or numerical) and the possibility that various assessments relate to the same consideration, the assessments cannot be converted immediately to variables. If the assessments are related to the same concept, they are converged into the same variable, while assessments with various values are either divided into two variables, which either relate to the minimum and maximum of the observed values (if the data is numerical) or contemplate all the assessments values into variables by being encoded through the use of dummy variables (in the case of categorical data) [19].

Lastly, the procedure of mapping raw EHR data is automated in order to obtain a data matrix representation, due to the high quantity of data and variables. To do so, a routine that takes the raw data retrived from the EHRs is applied, which produces as output results a data matrix representation of the database in which, each column symbolizes a variable and each line an patient episode.

4.2 Filter

A mRMR filter is chosen [18] to retrieve only the most important information from the database. That selection is made because the mRMR filter is well documented amongst the literature [18] and due to the success of its implementation in other studies [20] [21].

The mRMR filter is defined by both the maximal relevance and minimal redundancy criterions, and it has the goal to find, amid a database, a group of variables (amongst all of the independent attributes of the database) that optimally define the class attribute (target classification variable). As the optimal characterization of the target class variable refers to a minimization of the classification error and requires a maximal statistical dependency of the class variable of the group of selected independent variables (maximal dependency). So as to obtain this maximum dependency, a maximal relevance attribute selection criterion is used, which selects the variables with the utmost relevance to the target class attribute. This relevance is then defined in terms of mutual information. In the maximal relevance method, the selected variables are required to individually have the biggest mutual information with the target class variable, making the group of variables the N most important individual attributes.

Nonetheless, by only using the maximal relevance criterion, it is possible that the selected variables have a great deal of redundancy, as the dependency between the selected attributes could be substantial. Furthermore, when two attributes are significantly depended on each other and one of them is removed, there is not a substantial reduction on the ability of the remaining group of variables to discriminate the target class attribute. This minimal redundancy condition is then used to, along with the maximal relevance criterion, select the group that includes the most important individual attributes. Due to the minimal redundancy criterion, attributes that contain very similar and redundant information can be excluded.

4.3 Constructed datasets

When applied to a database, the mRMR filter selects the independent variables (in that database) that it considers to be the most appropriate for the prediction process of a chosen binary class variable. Taking this into account, from the database, the independent variables that relate the most to a specific group of DRGs can be selected (the relation can be to a MDC but also to DRGs of various MDCs), through the use of a binary class variable that represents, for each of the instances in the database, the presence or absence of their DRG in the selected group of DRGs. Therefore, through the use of the mRMR filter, particular datasets for specific groups of DRGs can be constructed.

In terms of datasets, the MDC 4 dataset and the MDC 5+1+16 dataset are chosen. The MDC 4 dataset is based on DRGs of the 'Diseases and disorders of the respiratory system' MDC (MDC 4), and it is constructed by considering as the binary class variable for the mRMR filter the presence (or absence) of each instance's DRG in the MDC 4 (Diseases and disorders of the respiratory system) and by selecting from the database the independent variables that relate the most with that binary categorical variable (MDC 4).

On the other hand, the MDCs 5+1+16 dataset is based on DRGs of the 'Diseases and disorders of the circulatory system' MDC (MDC 5) and on other disorders related to blood and the circulatory system that are present in some specific DRGs of the MDC 16 (Diseases and disorders of the blood, hematopoietic organs and immunological diseases) and MDC 1 (Diseases and disorders of the nervous system). In this case, the dataset is constructed by considering as the binary class variable for the mRMR filter the presence (or absence) of each instance's DRG in the MDC 5 (Diseases and disorders of the circulatory system) or in disorders related to blood and to the circulatory system of the MDC 16 and MDC 1. Being the mRMR filter used to select from the database, the independent variables that relate the most with the chosen binary categorical variable.

In both datasets the same observations are present, but due to the different applications of the mRMR filter [18] in each of the datasets, different independent variables are present. Thus, the goal of having various datasets can be defined as the opportunity to use in each one, only the most relevant variables for their case, while at the same time using all of the database observations in both datasets.

4.4 Regression models

4.4.1 Logistic regression

Since the prediction process is done based on a database of observations in which the corresponding DRG is known, a supervised learning methodology is applied. More precisely, it is applied the logistic regression to generate the prediction models for the DRGs through the support of the software Stata [22], a method in which a binary variable is used to represent the presence or absence of each of the DRGs.

A logistic regression algorithm consists in a statistical technique that aims at finding the most fitting classification model that describes the relation between a set of independent variables and a target dependent variable (DRG) [23]. It has a binary outcome for the class variable (DRG) where there are only possible the values of 1 for the presence of a certain condition, and 0 for the absence of such condition. Furthermore, the algorithm does not predict the outcome directly, but rather the probability of the outcome [24]. Thus logistic regression, by taking into account the dataset variables, predicts the probability that a instance with certain known independent variables has of belonging to a certain class (presence of DRG) [25].

4.4.2 Proposed experiments

For both of the constructed datasets is proposed an assembly of different experiments, in which, through the use of a logistic regression machine learning technique, classification models are built for the prediction (of the presence or absence) of each DRG. Initially, it is considered an experiment regarding all of the independent variables present in the dataset, which is performed in order to have an overview of the predictive capability of models in which all of the diverse independent variables are present.

Furthermore, experiments that are based on the data blocks of the independent variables in the dataset are also proposed in order to explore the predictive potential of the ICD codes (diagnosis data) data blocks. Those are divided into two types of experiments, one in which the DRGs are predicted by only using the variables that refer to ICD codes, and another that takes into account the various data blocks of predictors that can be present in the dataset with the exception of ICD codes. Thus by excluding the ICD codes, the experiment uses only independent variables related to demographic data, medication, prescriptions, assessments, patient's personal history and allergies.

Subsequently, two experiments using stepwise logistic regression models are taken into account. One is done by starting with the lowest p-value variable, and by adding variables in accordance with their progressively larger p-value. The other one is done by starting with the most relevant variable for the mRMR filter and by adding variables in decreased order of relevance. These two experiments are performed in order to analyze the differences between the models generated with the mRMR filter and the p-value criterions in terms of the evaluation metrics. Additionally, in both experiments the relation between the amount of predictor variables in the models and their predictive capacity is to be studied through the use of the evaluation metrics.

4.5 Evaluation metrics

In order to analyze the performance of the models in the devised experiments, a variety of evaluation metrics are chosen. In terms of performance measures for models, the measures of sensitivity and predictive positive value (PPV) are used. Furthermore, the McFadden pseudo r-squared is applied as a measure of goodness-of-fit, in order to assert the extent to which the model predictors explain the variations in the class variable outcome. Finally, the coefficients and p-value are used in order to analyze the contributions of the statistically significant independent variables (predictors) in the models.

5. Results

In this section are presented the results for the application of the previous methodology. The database used in the study consisted of 5089 observations of inpatient episodes, collected in the year of 2013 in the Hospital Fernando Fonseca [26], which had been formerly treated in a previous work [19]. These episodes correspond to a discretization of data from patients' EHRs. Each of the 5089 observations of the database contained an assigned DRG that belonged to a specific MDC, being that the distribution of the inpatient episodes amongst the MDCs differed vastly depending on the MDC.

As seen in many applications of machine learning techniques to real world data, the chosen datasets presented an imbalance of the target class variable (the chosen DRG). So, in order to diminish the possible effects of the imbalance of the data in the selected datasets, only the DRGs that had in each dataset over 50 observations were selected for an analysis. In the MDC 4 dataset, five DRGs were chosen, while in the MDCs 5+1+16 dataset there were chosen seven DRGs.

In this section, the various proposed experiments are studied. After performing all of the chosen experiments, it is selected for each DRG, the model that presents the best value for the combination (product) of the sensitivity and PPV performance measures. In table 1, it is possible to observe the models that had the best performance in the MDCs 5+1+16, alongside the values of the pseudo r-squared, the performance measures of sensitivity, PPV and their product, and finally the description of the DRGs. In table 2 it is possible to observe the models that had the best performance in the MDC 4 dataset. Furthermore, in each of these models, the most significant variables (p-value lower than 0.05) and the relation of their coefficients with the predicted DRG is analyzed.

Table 1. Models with the best performance measures for the DRGs using the MDCs 5+1+16 dataset.

DRG	Type of Model	P. R^2	Sens.	PPV	Sens. * PPV	DRG description
127	Only ICD codes	0,399	84,76%	33,15%	28,09%	Heart failure and / or shock
544	39 variables selected by the p-value criterion	0,375	52,11%	31,31%	16,32%	Congestive heart failure and / or cardiac arrhythmia, with major CC
138	41 variables selected by the p-value criterion	0,495	60,83%	30,13%	18,33%	Arrhythmia and / or cardiac conduction disorders, with CC
139	45 variables selected by the mRMR filter criterion	0,641	62,45%	52,70%	32,91%	Arrhythmia and / or cardiac conduction disorders without CC
395	50 variables selected by the p-value criterion	0,331	26,30%	34,41%	9,05%	Disruption of erythrocytes, age> 17 years
14	16 variables selected by the p-value criterion	0,581	77,76%	50,92%	39,60%	Cerebrovascular accident with infarction
15	5 variables selected by the p-value criterion	0,496	79,44%	27,35%	21,72%	Unspecific cerebrovascular accident and / or pre-cerebral occlusion without infarction

Table 2. Models with the best performance measures for the DRGs using the MDC 4 dataset.

DRG	Type of Model	P. R^2	Sens.	PPV	Sens. * PPV	DRG description
541	16 variables selected by the p-value criterion	0,343	67,88%	37,98%	25,78%	Respiratory disorders, except infections, bronchitis or asthma, with major CC
89	16 variables selected by the p-value criterion	0,45	81,76%	41,63%	34,04%	Pneumonia and / or simple pleurisy, age> 17 years, with CC
96	22 variables selected by the p-value criterion	0,444	46,46%	41,24%	19,16%	Bronchitis and / or asthma, age> 17 years, with CC
90	24 variables selected by the p-value criterion	0,448	40,53%	17,22%	6,98%	Pneumonia and / or simple pleurisy, age> 17 years without CC
87	48 variables selected by the mRMR filter criterion	0,303	36,59%	7,49%	2,74%	Pulmonary edema and / or respiratory failure

6. Discussion

The discussion of the results for the models that, amongst all the experiments, provide the best values for the combination of the performance measures in each of the DRGs, as well as the relation between the most significant variables with those DRG models, is performed in this section.

Regarding the most preponderant models, it was observed that there was an overall preponderance of the p-value criterion, with 9 out of the 12 DRGs having the best model through it, while only two excelled through the use of the RMR filter, and only one had its best model using only ICD codes. Thus higher performance of the p-value criterion models must be related with a more well-ordered choice of the most important variables for the DRGs prediction models.

Furthermore, in both the measures of sensitivity and PPV, it can be seen that the performance of the models varies greatly amongst DRGs (even on DRGs of the same dataset), thus leading to the conclusion that the same methods applied to different DRGs, even when using the same dataset, can conduce to very distinct results.

In terms of the relation of the performance measures with the imbalanced data in the datasets, it can be observed that in the MDC 4 dataset the DRG models that have the worst performances are also the models that present the lowest number of positive examples. Nonetheless, this phenomenon does not occur in the MDC 5+1+16 dataset, were the model with the lowest performance has more observations than 3 other variables that outperform this DRG model. Additionally, in neither of the datasets are the DRG models with the best performances those with the most observations. Therefore, these results can indicate that even though it is necessary a certain number of positive examples in the dataset, it is not the most important issue for the performance of the models, being the performance more related with the information present in the dataset variables and with the characteristics that each DRG has.

Regarding the measures of sensitivity and PPV, it can also be seen that there is a preponderance of the sensitivity in all of the DRG models, as the sensitivity is higher than the PPV in every DRG model, being sometimes much higher than the PPV. Furthermore, 8 out of the 12 DRGs studied have a model with a sensitivity marker above 50%, while only 2 DRGs (DRG 14 and DRG 139) have models with a PPV value over 50 %. This means that, although for most DRGs the produced models have a high level of certainty, they will not miss observations which present the DRG, there is a low certainty level that the models are not incorrectly classifying observations that do not present the DRG as positive.

In terms of goodness-of-fit, it can be observed that the models for the DRGs that have the lowest pseudo r-squared values also have the lowest values for the performance measures. Additionally, the highest pseudo r-squared values corresponded to the DRG models with the highest performance measures. This was expected, since the higher the pseudo r-squared value, the bigger is the quantity of the variation of the output class variable (DRG). This is explained by the independent variables (predictors), thus it is normal for models with high performance measures to also have high value of pseudo r-squared.

Concerning the evaluation of the most significant variables of the models, it can be stated that there is a great preponderance of ICD codes followed by medication, while prescriptions and assessments constitute a small number. Furthermore, the models that have the best results are those that have ICD code variables which relate directly with the DRGs. This follows the expectations due to the influence of ICD codes in the characterization of DRGs.

It is also possible to conclude in the analysis of most significant variables of the DRG models, that the variables with biggest positive coefficient values, are typically the variables that relate the most with the DRG. On the other hand, variables with negative coefficients generally do not have a relation with the DRG, thus decrease the probability of prediction the DRG. Nonetheless, there are some variables which are significant and present positive coefficients, while at the same time opposing the delimited DRG. This occurs in the two DRGs (DRG 541 and DRG 15) with more vast and not well defined groups, thus making the models more prone to the incorrect use of the independent variables.

7. Conclusion

In this work, was proposed a methodology for the prediction of DRGs, an important tool to the financing of hospital institutions. The process was performed through the use of structured data from EHRs and a logistic regression machine learning technique to build the DRG prediction models. Concerning what was assessed by the literature review, no other studies predicting DRGs from structured data of EHRs were found, so this one tries to give an insight into the theme.

Regarding the best models for the different DRGs, a great difference in terms of performance measures of the models amongst the DRGs was seen, leading to the conclusion that equal methodologies in different DRGs gave different results due to the diversity and complexity of the various DRGs.

As addressed in the discussion, in terms of the models performance measures, the sensitivity measure was constantly higher than the PPV measure. Additionally, the sensitivity measure had a typically good performance (over 50%) for most of the DRG models, while the PPV had relatively low values for most of the DRG models. This type of results, specifically concerning the high values of sensitivity, characterize the DRG models as less likely to miss a DRG, and thus help the hospital institution not to lose revenues. Oppositely, regarding the low values of the PPV measure, this may lead the models to predict a DRG incorrectly, which could lead the hospital to incur in fines for wrongly predicting a DRG. Thus, these models could be a basis, that with some improvement could be applied to a further DRG prediction system, which had an emphasis in not missing DRGs.

In conclusion, this study indicates that it is possible to predict DRGs with the use of structured data from EHRs with a certain degree of confidence. Nonetheless, as the results and their discussion have stated, this is still a work in progress that is in need of some improvements in order to enhance its performance and applicability to more DRGs.

References

[1] ACSS, “Grupos de Diagnósticos Homogéneos,” 2015. [Online]. Available: http://www.acss.min-saude.pt. [Accessed: 17-Mar-2015].

[2] P. P. Barros and G. Braun, “Upcoding in a NHS ∗,” pp. 1–27, 2012.

[3] C. Costa, S. Lopes, and R. Santana, “Diagnosis Related Groups e Disease Staging: importância para a administração hospitalar,” *Rev. Port. Saúde Pública*, pp. 7–28, 2008.

[4] R. Santana, “O financiamento hospitalar e a definição de preços,” *Rev. Port. Saúde Pública*, vol. 5, pp. 93–118, 2006.

[5] C. M. Borges, R. Ramalho, M. Bajanca, T. Oliveira, M. J. Major, P. Diz, and V. Rodrigues, “Implementação de um sistema de custeio por actividades nos hospitais do SNS,” pp. 141–160, 2010.

[6] P. Gottlober, “Medicare hospital prospective payment system: How DRG rates are calculated and updated,” 2001.

[7] R. Busse, A. Geissler, A. Aaviksoo, F. Cots, U. Hakkinen, C. Kobel, C. Mateus, Z. Or, J. O’Reilly, L. Serden, A. Street, S. S. Tan, and W. Quentin, *Diagnosis related groups in Europe: moving towards transparency, efficiency, and quality in hospitals*, vol. 346. 2013.

[8] R. F. Averill, N. Goldfield, J. S. Hughes, J. Bonazelli, E. C. McCullough, B. a Steinbeck, R. Mullin, A. M. Tang, J. Muldoon, L. Turner, and J. Gay, “All Patient Refined Diagnosis Related Groups (APR-DRGs): Methodology Overview,” 2003.

[9] S. Duckett, “The Development of Australian Refined Diagnosis Related Groups: the Australian Inpatient Casemix Classification,” *CaseMix Q.*, vol. 2, no. 4, pp. 115–120, 2000.

[10] M. L. Müller, T. Bürkle, S. Irps, N. Roeder, and H.-U. U. Prokosch, “The diagnosis related groups enhanced electronic medical record,” *Int. J. Med. Inform.*, vol. 70, no. 2–3, pp. 221–228, 2003.

[11] W. Quentin, A. Geissler, D. Scheller-kreinsen, and R. Busse, “DRG-type hospital payment in Germany: The G-DRG system,” *Euro Obs.*, vol. 12, pp. 4–6, 2002.

[12] M. M. Bellanger and L. Tardif, “Accounting and reimbursement schemes for inpatient care in France,” *Health Care Manag. Sci.*, vol. 9, no. 3, pp. 295–305, Jul. 2006.

[13] A. Lowe, “Casemix accounting systems and medical coding,” *J. Organ. Chang. Manag.*, vol. 14, no. 1, pp. 79–100, Feb. 2001.

[14] P. Clarke, J. Leal, C. Kelman, M. Smith, and S. Colagiuri, “Estimating the Cost of Complications of Diabetes in Australia Using Administrative Health-Care Data,” *Value Heal.*, vol. 11, no. 2, 2008.

[15] K. Okamoto, T. T. Uchiyama, and T. Takemura, “Automatic selection of diagnosis procedure combination codes based on partial treatment data relative to the number of hospitalization days,” in *APAMI 2012 Conference Proceedings*, 2012, vol. XXXIII, no. 2, pp. 81–87.

[16] T. Viangteeravat, “Potential identification of pediatric asthma patients within pediatric research database using low rank matrix decomposition,” pp. 1–8, 2013.

[17] J. Banks, “STATA,” *Econ. J.*, vol. 102, no. 415, p. 1581, 1992.

[18] H. Peng, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance , and Min-Redundancy,” *Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[19] J. C. Ferrão, F. Janela, M. D. Oliveira, and H. M. G. Martins, “Using structured EHR data and SVM to support ICD-9-CM coding,” *Proc. - 2013 IEEE Int. Conf. Healthc. Informatics, ICHI 2013*, pp. 511–516, 2013.

[20] P. A. Mundra and J. C. Rajapakse, “SVM-RFE with MRMR filter for gene selection,” *IEEE Trans. Nanobioscience*, vol. 9, pp. 31–37, 2010.

[21] A. El Akadi, A. Amine, A. El Ouardighi, and D. Aboutajdine, “A two-stage gene selection scheme utilizing MRMR filter and GA wrapper,” *Knowl. Inf. Syst.*, vol. 26, no. 3, pp. 487–500, 2011.

[22] StataCorp, “STATA User’s Guide: Release 14,” Stata Press, 2015.

[23] S. Rabe-Hesketh and B. Everitt, *A Handbook of Statistical Analyses using Stata*, Third. 2003.

[24] J. S. Long and J. Freese, *Regression models for categorical dependent variables using Stata*. 2006.

[25] R. P. Burns and R. Burns, *Business Research Methods and Statistics Using SPSS*. 2008.

[26] Hospital Prof. Doutor Fernando Fonseca, “Relatório e Contas,” Amadora, 2013.